Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

➔ some of the inferences on the analysis of the categorical variables and their effect on the dependent variable.

1. The season of Fall has the highest median followed by summer as they have the best weather conditions.
2. The median bike rentals have increased in the year 2019 compared to the year 2018. This may be due to the people getting conscious about the environment.
3. The bike rentals are more on non-holiday days compared to holiday. This indicates that people prefer to spend time at home during the holidays.
4. The months of Fall - June to October have a higher median value.
5. The overall median for the weekdays and working-days are the same.
6. The Clear weather situation has the highest median while the weather situation of Light snow has the least. The count of bike sharing is Zero for the weather Situation-4 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog'.

2. Why is it important to use drop_first=True during dummy variable creation?

➔ Dummy variables will be correlated if you don't remove the first column (redundant). This may have a negative impact on some models, and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importances may be distorted. Another argument is that having all dummy variables results in multicollinearity between them. We lose one column to keep everything under control.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

➔ The numerical variable 'atemp' has the highest correlation with the target variable 'cnt' with a value of '0.65' followed by 'temp' with a value of '0.64'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

➔ Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variabthe.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

➔ The top 3 features contributing significantly towards explaining the demands of the shared bikes:

1. temp(Temperature) - A coefficient value of '0.5480' indicates that a unit increase in temp variable, increases the bike hire numbers by 0.5480 units.
2. Light Snow(weathersit) A coefficient value of '-0.2838' indicates that, a unit increase of this variable, decreases the bike hire numbers by -0.2838 units.
3. Yr(Year) - A coefficient value of '0.2328' indicates that, a unit increase of this variable, increase the bike hire numbers by 0.2328 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

➔ Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables. There are two types of linear regression- simple linear regression and multiple linear regression. Simple linear regression is used when a single independent variable is used to predict the value of the target variabthe. Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

2. Explain the Anscombe’s quartet in detail.

➔ A regression model is not always necessarily an exact one, it can also be fooled by some (smart) data! In certain cases, there are multiple datasets which are completely different but after training, the regression model looks the same. A group of four such datasets having identical descriptive statistics but with some peculiarities, is the Anscombe's quartet.

3. What is Pearson’s R?

➔ Pearson's r is a numerical representation of the strength of the linear relationship between the variables. Its value ranges from 1 to +1. It depicts the linear relationship of two sets of data. In layman's terms, it asks if we can draw a line graph to represent the data.

r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

➔ Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitusome and oneits. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model. The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

➔ **VIF - the variance inflation factor** :- The VIF indicates how much collinearity has increased the variance of the coefficient estimate. (VIF) is equal to $1/(1 - Ri\text{^}2)$. VIF = infinity if there is perfect correlation. Where R-1 denotes the R-square value of the independent variable for which we want to see how well it is explained by other independent variables. If an independent variable can be completely described by other independent variables, it has perfect correlation and has an R-squared value of 1. As a result, $VIF = 1/(1 - 1)$ provides $VIF = 1/0$, which is "infinity."

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

➔ The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.